A penalized likelihood approach for a progressive three-state model with censored and truncated data: application to AIDS.

Pierre Joly, Daniel Commenges

A penalized likelihood approach for a progressive three-state model with censored and truncated data: Application to AIDS

Pierre JOLY, Daniel COMMENGES

INSERM U. 330

Université de Bordeaux II

146, rue Léo Saignat

33076 Bordeaux Cedex, France

Tel : (33) 5 57 57 11 36

Fax : (33) 5 56 99 13 60

E-mail: pjoly@u-bordeaux2.fr

Corresponding author: Pierre JOLY

URL: www.isped.u-bordeaux2.fr/isped/recherche/biostats/fr-biostats-accueil.htm

October 29, 1998

Summary

We consider the estimation of the intensity and survival functions for a continuous time progressive three-state semi-Markov model with intermittently observed data. The estimator of the intensity function is defined non-parametrically as the maximum of a penalized likelihood. We thus obtain smooth estimates of the intensity and survival functions. This approach can accommodate complex observation schemes such as truncation and interval censoring. The method is illustrated with a study of hemophiliacs infected by HIV. The intensity functions and the cumulative distribution functions for the time to infection and for the time to AIDS are estimated. Covariates can easily be incorporated into the model.

Key words: Three-state semi-Markov model, intensity function, penalized likelihood, splines, truncation, interval-censoring.

1 Introduction

The three state model is useful in a variety of biomedical settings, especially those concerned with characterizing an individual's progression through various stages of a disease. Many authors have worked on methods for analyzing the three-state model in the context of AIDS (De Gruttola and Lagakos, 1989; Bacchetti and Jewell, 1991; Frydman, 1992; Frydman, 1995; Kim, De Gruttola and Lagakos, 1993) and carcinogenicity testing (McKnight and Crowley, 1984; Lindsey and Ryan, 1993; Kodell and Nelson, 1980; Portier and Dinse, 1987). Analysis in both these contexts is complicated by the fact that the disease process is observed only intermittently. None of the currently proposed methods are entirely satisfactory when it comes to estimating the hazard or intensity function. The non-parametric approaches tend to be unstable, while parametric approaches impose too many assumptions. In this paper, we use smoothing methods to provide a compromise between these two extremes.

Our approach is based on the penalized likelihood. We introduce an *a priori* knowledge of smoothness of the intensity functions, by penalizing the log-likelihood by the sum of the square norms of the second derivative of the intensity functions. The estimators are defined non-parametrically as the functions which maximize the penalized likelihood. The solution is then approximated using splines. This approach, presented in section 2, is an extension to a three-state model of methods for survival analysis, proposed by O'Sullivan (1988) and Joly, Commenges and Letenneur (1998). We also show how it can be applied to regression models including the proportional hazards model. In section 3, the proposed method is applied to AIDS data taken from De Gruttola and Lagakos (1989).

2 The Model

2.1 The three-state model

The three-state model with irreversible transition is depicted in Figure 1. In its application to AIDS, state 0 is "uninfected", state 1 is "infected" and state 2 is "AIDS". The dates of transition between states may not be known exactly. When there are successive visits, the time of occurrence of the event of interest is only known to lie between two visits. In the case considered in this paper, the time spent in state 0 may be left-truncated, interval-censored or right-censored and the time spent in state 1 may be right-censored. Thereafter, F_{lk} is the cumulative distribution function, α_{lk} is the intensity function and A_{lk} is the cumulative intensity function associated with the time spent in state l, $l = 0, 1$; for the progressive model considered, $k = l + 1$.

2.2 Observations and likelihood

Let X_i^0 be the time spent in state 0 by subject i. $\mathcal{L}_i^0$ is a fixed truncation time. X_i^0 is left-truncated because only individuals with $X_i^0 > \mathcal{L}_i^0$ are observed. X_i^0 is interval-censored, i.e., we only know that it lies in a known interval $[L_i^0, R_i^0]$. Given the previous observations of the process, we assume that the conditional intensity of the point process of examination times is independent of the process of transitions (examinations do not occur in response to the state of the patient).

Let X_i^1 be the time spent in state 1. We assume that X^0 and X^1 are independent (this is a semi-Markov model). T_i is the time the subject i was last seen. Therefore, T_i may be the time of right censoring for the first transition (in this case $T_i = L_i^0$), the time of right censoring for the second transition (in this case $T_i = L_i^1$) or the time of the second transition. For the latter case, we have $X_i^1 = T_i - X_i^0$ with X_i^0 not exactly known. If X_i^0 is right-censored, X_i^1 is not observed at all. We assume that the censoring time L_i^1, for the second transition,

is independent of X_i^1.

The log-likelihood, conditional on the event $X_i^0 > \mathcal{L}_i^0$, for n independent observations, is

$$l = \sum_{i=1}^{n} \log \left\{ \frac{1}{e^{-A_{01}(\mathcal{L}_i^0)}} \int_{L_i^0}^{R_i^0} e^{-A_{01}(u)} \left(\alpha_{01}(u) e^{-A_{12}(T_i-u)}\right)^{\delta_{1i}} \left(\alpha_{12}(T_i-u)\right)^{\delta_{2i}} du \right\} \quad (1)$$

where the indicator δ_{1i} is equal to 0 if subject i is right-censored for the first transition ($R_i^0 = +\infty$), and is 1 otherwise. Similarly, the indicator δ_{2i} is defined for the second transition; as noted above, if $\delta_{1i} = 0$ then $\delta_{2i} = 0$. We shall denote the log-likelihood as $l(\alpha_{01}, \alpha_{12})$, since it can be expressed as a function of $\alpha_{01}(.)$ and $\alpha_{02}(.)$.

In the application in section 3, the data do not present left-truncation, thus $e^{-A_{01}(\mathcal{L}_i^0)} = 1$.

2.3 Penalized likelihood

In real life smooth intensity functions may be expected. To introduce such *a priori* knowledge, we penalize the likelihood by a term which takes large values for rough functions. The roughness penalty function chosen for the three-state model is the sum of the square norms of the second derivatives of the intensities. The penalized log-likelihood is thus defined as

$$pl(\alpha_{01}, \alpha_{12}) = l(\alpha_{01}, \alpha_{12}) - \kappa_1 \int \alpha_{01}^{''\,2}(u) du - \kappa_2 \int \alpha_{12}^{''\,2}(u) du \quad (2)$$

where κ_1 and κ_2 are two positive smoothing parameters which control the trade-off between the data fit and the smoothness of the functions. Maximization of (2) defines the maximum penalized likelihood estimators (MPnLE) $\hat{\alpha}_{01}(.)$ and $\hat{\alpha}_{12}(.)$ and hence $\hat{F}_{01}(.)$ and $\hat{F}_{12}(.)$. The smoothing parameters are chosen by the cross-validation method presented in Joly et al. (1998). To choose the smoothing parameters, and only for this step, we separate the three-state model into two survival models. To choose κ_1 we use the approximate cross-validation score for

a survival model with left-truncated and interval- or right-censored observations. For the second transition, the exact date of infection is imputed to be the middle of the censoring interval. κ_2 is chosen using the approximate cross-validation score for a survival model with only right-censored observations. Such an approach can be used because the estimators are not very sensitive to small variations of the smoothing parameters. It is theoretically possible to extend the cross-validation method to the three-state model considered, but the maximum of the approximate cross-validation score in two dimensions is difficult to obtain numerically.

2.4 Approximation of the estimators

The MPnLE cannot be calculated explicitly. However, it can be approximated to any degree of accuracy using splines. Splines are piecewise polynomial functions which are combined linearly to approximate a function on an interval. We use M-splines and I-splines, which are a variants of B-splines. For more details, see Ramsay (1988) and Joly et al. (1998).

The estimator $\hat{A}(.)$ for a given transition is approximated by a linear combination of m I-splines: $\tilde{A}(.) = \tilde{\boldsymbol{\theta}}\boldsymbol{I}(.)$, where $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, ..., \tilde{\theta}_m)$ and $\boldsymbol{I}(.) = (I_1(.), ..., I_m(.))^T$. By differentiation we obtain: $\tilde{\alpha}(.) = \tilde{\boldsymbol{\theta}}\boldsymbol{M}(.)$, where $\boldsymbol{M}(.) = (M_1(.), ..., M_m(.))^T$. As there are two distinct times, we use two distinct bases of splines, possibly with a different number of splines in each basis.

The approximation $\tilde{\alpha}$ of $\hat{\alpha}$ is the function belonging to the space generated by the basis of splines, which maximizes $pl(\alpha_{01}, \alpha_{12})$. The general penalized log-likelihood for our case is therefore

$$\sum_{i=1}^{n} \log \left(\frac{1}{e^{-\boldsymbol{\theta}_1 \boldsymbol{I}_1(\mathcal{L}_i^0)}} \int_{L_i^0}^{R_i^0} e^{-\boldsymbol{\theta}_1 \boldsymbol{I}_1(u)} \left\{ (\boldsymbol{\theta}_1 \boldsymbol{M}_1(u)) e^{-\boldsymbol{\theta}_2 \boldsymbol{I}_2(T_i - u)} \right\}^{\delta_{1i}} (\boldsymbol{\theta}_2 \boldsymbol{M}_2(T_i - u))^{\delta_{2i}} \, du \right)$$

$$-\kappa_1 \int (\boldsymbol{\theta}_1 \boldsymbol{M}_1''(u))^2 \, du - \kappa_2 \int (\boldsymbol{\theta}_2 \boldsymbol{M}_2''(u))^2 \, du.$$

The vectors $\tilde{\boldsymbol{\theta}}_1$ and $\tilde{\boldsymbol{\theta}}_2$ for fixed κ_1 and κ_2 are obtained by maximizing the log-likelihood using a Marquardt's algorithm (1963), which is a compromise between the steepest descent and Newton-Raphson algorithms. When the two vectors $\tilde{\boldsymbol{\theta}}_1$ and $\tilde{\boldsymbol{\theta}}_2$ are obtained, all the functions of interest can be computed, as in a parametric method.

2.5 Extension to regression models

As in Joly et al. (1998) the penalized likelihood can be used to estimate the intensity functions in a general regression model defined by: $\alpha_{lk}^i(.) = \varphi[\alpha_{lk}^0(.), \boldsymbol{z}_{lk}^i \boldsymbol{\beta}_{lk}]$, where $\alpha_{lk}^0(.)$ is the baseline intensity function from state l to state k, $\boldsymbol{\beta}_{lk}$ is a vector of regression parameters and $\boldsymbol{z}_{lk}^i$ is the vector of covariates for subject i. Note that the proportional hazards model is obtained by choosing $\varphi[u(.), v] = u(.)e^v$. The penalized log-likelihood used is therefore

$$l[\alpha_{01}^0(.), \alpha_{12}^0(.), \boldsymbol{Z}_{01}\boldsymbol{\beta}_{01}, \boldsymbol{Z}_{12}\boldsymbol{\beta}_{12}] - \kappa_1 \int \alpha_{01}^{0\prime\prime 2}(u)du - \kappa_2 \int \alpha_{12}^{0\prime\prime 2}(u)du \qquad (3)$$

where $\boldsymbol{Z}_{lk}$ is the matrix with rows equal to $\boldsymbol{z}_{lk}^i$, $i = 1, ..., n$. The regression parameters and the baseline functions are estimated simultaneously by Marquardt's method.

2.6 Confidence intervals

The confidence intervals may be determined using the nonparametric percentile bootstrap technique (Hall, 1992; Wang and Wahba, 1995)

To evaluate the method, we performed a simulation study. Specifically, we applied the following steps: the data are resampled 200 times. For each point considered, we order the 200 values obtained; the lower bound and the upper bound are respectively given by the 2.5th and the 97.5th empirical percentiles. Samples of size 100 were generated from a gamma distribution. For each of 500 replications, we generated random samples $X_1^0, ..., X_n^0$ and $X_1^1, ..., X_n^1$ of i.i.d failure times

and $C_1^0, ..., C_n^0$ and $C_1^1, ..., C_n^1$ of i.i.d censoring times, where the C_i were independent of the X_i. The observed samples were $(Y_1^0, \delta_{11}, Y_1^1, \delta_{21}), ..., (Y_n^0, \delta_{1n}, Y_n^1, \delta_{2n})$ where $Y_i = \min(X_i, C_i)$ and $\delta_i = \mathrm{I}_{[X_i \leq C_i]}$. Note that if $\delta_{1i} = 0$ we ignored Y_i^1 and δ_{2i}. X^0, X^1, C^0 and C^1 have Gamma distributions $(\alpha^\gamma t^{\gamma-1} e^{-\alpha t}/\Gamma(\gamma))$ with parameters $(\gamma; \alpha)$: (15;2), (15;2), (20;2) and (15;2), respectively. The percentage of censoring was around 20% for the first transition and 50% for the second. The smoothing parameters were chosen as explained above only for the first replication and were kept constant for the subsequent ones. The number of knots was set to 7. The coverage rates of bootstrap confidence intervals of the survival function at the four quintiles for the two transitions are given in Table 1, which suggests that the method worked reasonably well.

3 Application to AIDS

To illustrate the method, we present an application to modeling the risk of HIV infection and the risk of AIDS onset. The observations $(L_i^0, R_i^0, T_i, \delta_{2i})$ for 262 subjects are published in the articles of De Gruttola and Lagakos (1989) and Frydman (1992) (there is no left-truncation). There are two distinct groups of subjects. Among the 157 subjects of the "lightly treated" group, 95 had become infected and 14 of these had developed AIDS or other clinical symptoms. In the "heavily treated" group, there were 105 subjects, and among the 97 who became infected, 29 developed AIDS or other clinical symptoms.

We used 12 knots and M-splines of order 4 for the approximation of each intensity function. The smoothing parameters were chosen by the cross-validation method as described in section 2.3. Then we maximized (2) for the values of κ_1 and κ_2 obtained with this first step. The cumulative distribution functions of the two groups, presented in Figures 2 and 3, are in agreement with those of De Gruttola and Lagakos (1989). Confidence intervals were evaluated using the method decribed in section 2.6 applied to 100 equidistant points. Figures 4

and 5 display the estimated intensity functions for the two groups and the two transitions. It may be seen that the "heavily treated" group had a higher risk than the "lightly treated" group of being infected and developing AIDS.

The proportional hazards assumption is particularly easy to check visually by examining the graph of the smoothed estimated intensities. Indeed Figure 5 suggests that a proportional hazards assumption holds for the treatment for the second transition. If we perform a semi-parametric proportional hazards model, as described in section 2.5, the estimate of the relative risk to develop AIDS between the "heavily treated" group and the "lightly treated" group is 2.22 (95% confidence interval $[1.16, 4.24]$).

4 Discussion

We have shown that the penalized likelihood approach yields a method for analyzing data arising from complex observation schemes and for providing estimators of the intensity functions, which cannot be estimated using conventional non-parametric methods. The approach can be applied to non-homogeneous Markov models as well as to semi-Markov models. It should be possible to treat more complex multi-state models, although numerical problems may arise.

REFERENCES

Bacchetti, P. and Jewell, N. P. (1991). Nonparametric estimation of the incubation period of AIDS based on a prevalent cohort with unknown infection times. *Biometrics* **47,** 947-960.

De Gruttola, V. and Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics* **45,** 1-11.

Frydman, H. (1992). A non-parametric estimation procedure for a periodically observed three-state Markov process, with application to Aids. *Journal of the Royal Statistical Society, Series B* **54,** 853-866.

Frydman, H. (1995). Semi-parametric estimation in a three-state duration-dependent Markov model from interval-censored observations with application to Aids. *Biometrics* **51,** 502-511.

Hall, P. (1992). *The bootstrap and edgeworth expansion.* Springer-Verlag.

Joly, P., Commenges, D. and Letenneur, L. (1998). A penalized likelihood approach for arbitrarily censored and truncated data: application to age-specific incidence of dementia. *Biometrics*, **54,** 185-194.

Kim, M.Y., De Gruttola, V. and Lagakos, S. W. (1993). Analyzing doubly-censored data with covariates, with application to AIDS. *Biometrics* **49,** 13-22.

Kodell, R. L., and Nelson, C. J. (1980). An illness-death model for the study of the carcinogenic process using survival/sacrifice data. *Biometrics* **36,** 267-277.

Lindsey, J. C. and Ryan, L. M. (1993). A three-state multiplicative model for rodent tumorigenicity experiments. *Journal of the Royal Statistical Society, Series C* **42,** 283-300.

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal of Applied Mathematics* **11,** 431-441.

McKnight, B. and Crowley, J. (1984). Tests for differences in tumor incidence based on animal carcinogenesis experiments. *Journal of the American Statistical Association* **79,** 639-648.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing* **9,** 363-379.

Portier, C. J. and Dinse, G. E. (1987). Semiparametric analysis of tumor incidence rates in survival/sacrifice experiments. *Biometrics* **43,** 107-114.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science* **3,** 425-461.

Wang, Y. and Wahba, G. (1995). Bootstrap confidence intervals for smoothing splines and their comparison to bayesian confidence intervals. *Journal on Statistical Computing Simulation* **51,** 263-279.

Captions for Figures

Figure 1: Progressive three-state semi-Markov Model.
τ is the time since the first transition.

Figure 2: Estimated cumulative distribution function of times of HIV seroconversion (solid line) for heavily (upper curve) and lightly treated groups (lower curve) and 95% confidence intervals (dashed line).

Figure 3: Estimated cumulative distribution function of induction times between HIV seroconversion and onset of symptoms (solid line) for heavily (upper curve) and lightly treated groups (lower curve) and 95% confidence intervals (dashed line).

Figure 4: Estimated intensity function of times of HIV seroconversion for heavily (solid line) and lightly treated groups (dashed line).

Figure 5: Estimated intensity function of induction times between HIV seroconversion and onset of symptoms for heavily (solid line) and lightly treated groups (dashed line).

Table 1

Coverage rate of bootstrap confidence intervals of the survival function at the four quintiles for the two transitions.

	0.2	0.4	0.6	0.8
First transition	94.4	94	92.6	92.8
Second transition	96	94.4	94	92.6